D208–D210 *Nucleic Acids Research, 2004, Vol. 32, Database issue*
DOI: 10.1093/nar/gkh042

IMGT/3Dstructure-DB and IMGT/StructuralQuery, a database and a tool for immunoglobulin, T cell receptor and MHC structural data

Quentin Kaas[1], Manuel Ruiz[1] and Marie-Paule Lefranc[1,2,*]

[1]IMGT, the international ImMunoGeneTics information system®, Laboratoire d'ImmunoGénétique Moléculaire, LIGM, Institut de Génétique Humaine IGH, UPR CNRS 1142, 141 rue de la Cardonille, F-34396 Montpellier Cedex 5, France and [2]Institut Universitaire de France

Received June 27, 2003; Revised and Accepted September 18, 2003

ABSTRACT

IMGT/3Dstructure-DB and IMGT/Structural-Query are a novel 3D structure database and a new tool for immunological proteins. They are part of IMGT, the international ImMunoGenetics information system®, a high-quality integrated knowledge resource specializing in immunoglobulins (IG), T cell receptors (TR), major histocompatibility complex (MHC) and related proteins of the immune system (RPI) of human and other vertebrate species, which consists of databases, Web resources and interactive on-line tools. IMGT/3Dstructure-DB data are described according to the IMGT Scientific chart rules based on the IMGT-ONTOLOGY concepts. IMGT/3Dstructure-DB provides IMGT gene and allele identification of IG, TR and MHC proteins with known 3D structures, domain delimitations, amino acid positions according to the IMGT unique numbering and renumbered coordinate flat files. Moreover IMGT/3Dstructure-DB provides 2D graphical representations (or Collier de Perles) and results of contact analysis. The IMGT/StructuralQuery tool allows search of this database based on specific structural characteristics. IMGT/3Dstructure-DB and IMGT/StructuralQuery are freely available at http://imgt.cines.fr.

INTRODUCTION

The immunoglobulins (IG), T cell receptors (TR), major histocompatibility complex (MHC) and related proteins of the immune system (RPI) are proteins which are extensively studied at the structural level. However, owing to the complexity of their polymorphisms and, for the IG and TR, of their synthesis (1,2), a specialized 3D database was needed to expertly identify the genes and alleles encoding these proteins. The IMGT/3Dstructure-DB was implemented to provide a unique resource of expertise with detailed specific annotations on structural data of IG, TR, MHC and RPI, from human and other vertebrate species, extracted from the Protein Data Bank PDB (3). IMGT/3Dstructure-DB is part of IMGT, the international ImMunoGenetics information system® (4), and contains standardized information on the sequences, 2D structures (Colliers de Perles) (5) and 3D structures (with links to other structural databases). The IMGT/Structural query tool was implemented to query the database based on specific structural characteristics. Amino acid positions are according to the IMGT unique numbering (http://imgt.cines.fr) (6), described in the IMGT Scientific chart, which is based on the NUMEROTATION concept of IMGT-ONTOLOGY (7). Structural analysis and standardization of amino acid numbering and gene name will ease large-scale comparative studies of IG and TR folds and loops and of MHC grooves, and analysis of ligand receptor and domain interactions.

IMGT/3Dstructure-DB DATABASE

Statistics

The IMGT/3Dstructure-DB database manages 634 coordinate files which correspond to 422 different proteins (260 IG, 18 TR and 144 MHC). IG structures include 62 *Homo sapiens*, 169 *Mus musculus*, four *Camelus dromedarius*, eight *Rattus rattus*, one *Cricetinae* gen. sp. and 16 engineered proteins. TR structures include five *H.sapiens* and 13 *M.musculus* proteins. MHC structures include 61 *M.musculus*, 80 *H.sapiens* and three *Cricetinae* gen. sp. proteins. Two hundred and six different V genes and alleles were identified in V-DOMAINs: 185 IG (97 IGHV, 16 IGLV, 72 IGKV) and 21 TR (11 TRAV, 7 TRBV, 2 TRDV and 1 TRGV).

IMGT/3Dstructure-DB query

IMGT/3Dstructure-DB is queried through a user-friendly CGI interface. The user can search (i) by PDB code, protein name, (ii) by reference, (iii) by receptor description, or (iv) by selecting a group, a subgroup, a gene or a chain type, and a species. The user chooses to see an 'Overview' table (with a list of the IMGT/3Dstructure-DB entries in PDB code order) or 'Sequence details' of the G-DOMAIN, C-DOMAIN or

*To whom correspondence should be addressed at IMGT, the international ImMunoGeneTics information system®, Laboratoire d'ImmunoGénétique Moléculaire, LIGM, Institut de Génétique Humaine IGH, UPR CNRS 1142, 141 rue de la Cardonille, F-34396 Montpellier Cedex 5, France.
Tel: +33 4 99 61 99 65; Fax: +33 4 99 61 99 01; Email: lefranc@ligm.igh.cnrs.fr

V-DOMAIN [or any part of it: framework (FR) or complementarity determining region (CDR)].

IMGT/3Dstructure-DB results

Two displays are available for the IMGT/3Dstructure-DB results: 'Overview' or 'Sequence details'.

The 'Overview' results table provides the list of the IMGT/3Dstructure-DB entries displayed with the PDB code, IMGT protein names, IMGT receptor description, species, ligand(s), experimental technique, resolution and PDB release date. Each entry is detailed in an IMGT/3Dstructure-DB card, accessible by clicking the number in the first column. The IMGT/3Dstructure-DB card comprises:

(i) for each protein entry, a summary table [IMGT protein name, IMGT receptor description, type (IG, TR or MHC), ligand(s), species and chain ID];

(ii) for the whole coordinate file, links to general structural databases (PDB, MMDB, OCAS and PQS) and bibliographical database (PubMed), reference, experimental technique, resolution and PDB release date;

(iii) a link to the contact analysis results: 'Contacts between domains'. Atoms are considered to be in contact when no water molecule can occur between them. The atomic contacts are gathered at the domain–domain and residue–residue level, their type is identified (polar, hydrogen bond, Van der Waals) and they are classified in backbone–backbone, side chain–side chain and backbone–side chain contacts;

(iv) a link to the IMGT/3Dstructure-DB file renumbered according to IMGT unique numbering (6). The file can be displayed on-line or downloaded;

(v) a link to view the 3D structure with Rasmol (chains are colored by chain type);

(vi) a detailed description of the individual chains: chain ID, chain description, chain amino acid sequence with domain and region delimitations, link to sequence databases (Swiss-Prot, PIR, GenBank, etc.), characterization of each domain [domain description, IMGT gene and allele names (1,2), 2D graphical representation or Collier de Perles (5) sequence with IMGT gaps (6)]. For IG and TR V-DOMAINs, CDR-IMGT lengths and Collier de Perles on two layers with hydrogen bonds are also provided.

In the 'Sequence details' results page, chain ID, species and amino acid sequences of the selected domain are displayed with a link to the IMGT/3Dstructure-DB cards.

IMGT/StructuralQuery TOOL

The IMGT/StructuralQuery tools allows retrieval of IMGT/3Dstructure-DB entries using amino acid structural criteria: distance, ϕ and ψ angles, accessible surface area (SA), amino acid type (aa), CDR-IMGT lengths (6). Examples of queries are shown below. The mnemonics (distance, ϕ, ψ, SA, aa) are named, within parentheses, by the amino acid position according to IMGT unique numbering for V-DOMAIN, and the letter V (for V-DOMAIN).

Examples of query:

(i) distance (12V, 23V) < 12: allows the retrieval of all IMGT/3Dstructure-DB entries for which the alpha carbon of the amino acids at positions 12 and 23 in a V-DOMAIN are at a distance of <12 Å.

(ii) aa (2V) = S: allows the retrieval of all IMGT/3Dstructure-DB entries for which the amino acid at position 2 in a V-DOMAIN is a serine.

A structural query may use several criteria. In that case, criteria are coordinated by logical operators (AND or OR) and parentheses.

(iii) Length (CDR1) = 6 and [distance (12V, 23V) < 16 or distance (12V, 24V) < 16]: allows the retrieval of all IMGT/3Dstructure-DB entries comprising a V-DOMAIN with a CDR1-IMGT of six amino acids and for which the amino acid at position 12 is at a distance of <16 Å from positions 23 and 24 (between alpha carbons).

DATABASE AND TOOL IMPLEMENTATION

The IMGT/3Dstructure-DB database was implemented using MySQL. Programs were written in Perl. The file coordinates are extracted once a week from PDB (3) and selected by keywords checking through the file text. The program IMGT3D/AlleleAlign (8) was implemented for analysis of the amino acid sequences: IMGT gene and allele identification and region delimitation (V-REGION, J-REGION, etc.) are obtained by running sequentially the sequences with the FASTA program against the IMGT reference directory sets (IMGT Repertoire http://imgt.cines.fr). This program delimits the V-DOMAIN by combining the V-J-REGION or V-D-J-REGION, depending on the chain type. Amino acid IMGT numerotation is created by aligning the domain sequences with the IMGT reference sequences using a modified Smith and Waterman algorithm. IMGT/StructuralQuery uses a program to determine the distances between the α carbons of a domain, and the Stride program (9) to determine amino acid ϕ/ψ angle, accessible surface area and secondary structure. Chain partners are identified by combining gene and contact analysis.

CONCLUSION

IMGT/3Dstructure-DB integrates data from sequence and structural sources. This database provides, for the first time, the identification of IMGT genes and alleles expressed in the IG, TR and MHC with known 3D structures (5). This information is of high value since the IMGT gene names for IG and TR (1,2) was approved by the Human Genome (HUGO) Nomenclature Committee (HGNC) in 1999, and entered in LocusLink (NCBI) (10), Genew (11), GDB (12) and GeneCards (13). Moreover, IMGT/3Dstructure-DB provides also, for the first time, an identical numbering for positions in the 1D, 2D and 3D structures of antigen receptors, whatever the receptor type (IG or TR), the chain type (heavy, κ, λ for IG, and α, β, γ, δ for TR) and whatever the domain (V-DOMAIN or C-DOMAIN), and of RPI V-LIKE-DOMAINs or C-LIKE-DOMAINs. IMGT unique numbering has also been implemented for the G-DOMAIN of the MHC class I and class II whatever the species, and for the G-LIKE-DOMAINs of the RPI. IMGT/StructuralQuery is the first IMGT tool that makes extensive use of the IMGT unique numbering to make comparisons of structural data. These standardizations will be a great help in large-scale sequence–structure studies and more generally in protein engineering.

AVAILABILITY AND CITATION

Authors who use the IMGT/3Dstructure-DB and IMGT/StructuralQuery are encouraged to cite this article and to quote the IMGT Home page URL, http://imgt.cines.fr.

ACKNOWLEDGEMENTS

IMGT is funded by the European Union's 5th PCRDT programme (QLG2-2000-01287), the Centre National de la Recherche Scientifique (CNRS), the Ministère de la Recherche et de l'Education Nationale.

REFERENCES

1. Lefranc,M.-P. and Lefranc,G. (2001) *The Immunoglobulin FactsBook*. Academic Press, London, UK, 458 pp.
2. Lefranc,M.-P. and Lefranc,G (2001) *The T Cell Receptor FactsBook*. Academic Press, London, UK, 398 pp.
3. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
4. Lefranc,M.-P. (2003) IMGT, the international ImMunoGeneTics database. *Nucleic Acids Res.*, **31**, 307–310.
5. Ruiz,M. and Lefranc,M.-P. (2002) IMGT gene identification and Colliers de Perles of human immunoglobulins with known 3D structures. *Immunogenetics*, **53**, 857–883.
6. Lefranc,M.-P., Pommié,C., Ruiz,M., Giudicelli,V., Foulquier,E., Truong,L., Thouvenin-Contet,V. and Lefranc,G. (2003) IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. *Dev. Comp. Immunol.*, **27**, 55–77.
7. Giudicelli,V. and Lefranc,M.-P. (1999) Ontology for immunogenetics: IMGT-ONTOLOGY. *Bioinformatics*, **15**, 1047–1054.
8. Lefranc,M.-P., Giudicelli,V., Ginestoux,C., Bosc,N., Folch,G., Guiraudou,D., Jabado-Michaloud,J., Magris,S., Scaviner D., Thouvenin,V. *et al.* (2003) IMGT-ONTOLOGY for immunogenetics and immunoinformatics, http://imgt.cines.fr. *In Silico Biol.*, to be published online.
9. Frishman,D. and Argos,P. (1995) Knowledge-based secondary structure assignment. *Proteins*, **23**, 566–579.
10. Pruitt,K.D. and Maglott,D.R. (2001) RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.*, **29**, 137–140.
11. Wain,H.M, Lush,M., Ducluzeau,F. and Povey,S. (2002) Genew: the human gene nomenclature database. *Nucleic Acids Res.*, **30**, 169–171.
12. Letovsky,S.I., Cottingham,R.W., Porter,C.J. and Li,P.W. (1998) GDB: the Human Genome Database. *Nucleic Acids Res.*, **26**, 94–99.
13. Safran,M., Solomon,I., Shmueli,O., Lapidot,M., Shen-Orr,S., Adato,A., Ben-Dor,U., Esterman,N., Rosen,N., Peter,I. *et al.* (2002) GeneCards 2002: towards a complete, object-oriented, human gene compendium. *Bioinformatics*, **18**, 1542–1543.